Mitochondrial data are not suitable for resolving placental mammal phylogeny

Morgan, Claire C.; Creevey, Christopher J.; O'Connell, Mary J.

Published in:
Mammalian Genome

DOI:
10.1007/s00335-014-9544-9

Publication date:
2014

Citation for published version (APA):
Morgan, C. C., Creevey, C. J., & O'Connell, M. J. (2014). Mitochondrial data are not suitable for resolving placental mammal phylogeny. *Mammalian Genome*, *25*(11-12), 636-647. https://doi.org/10.1007/s00335-014-9544-9

Mitochondrial data are not suitable for resolving Placental mammal phylogeny

Claire C. Morgan [1,2,3], Christopher J. Creevey [4] and Mary J. O'Connell [1,2]*

[1]Bioinformatics and Molecular Evolution Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland.
[2]Centre for Scientific Computing & Complex Systems Modelling (SCI-SYM), Dublin City University, Glasnevin, Dublin 9, Ireland.
[3]Current address: National Heart and Lung Institute, Imperial College London, London W12 0NN, UK.
[4]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Wales, UK.

*Correspondence to be sent to: mary.oconnell@dcu.ie.
Bioinformatics and Molecular Evolution Group,
School of Biotechnology,
Dublin City University,
Glasnevin,
Dublin 9.
Ireland.
Phone: +353 1 700 5112

Keywords
Mammal, mitochondrial DNA, Data quality, Site-rate categorization, Site-Stripping, Phylogeny.

Abbreviations
AA: Amino Acids; BIC: Bayesian Information Criterion; bp: Base Pairs; CO1: Cytochrome c oxidase I; CYTB: Cytochrome b; LM: Likelihood Mapping; lnL: Natural Log of Likelihood; ML: Maximum Likelihood; MSA: Multiple Sequence Alignment; mtDNA: Mitochondrial DNA; mtGene: Mitochondrial Gene; nucDNA: Nuclear DNA; RF: Robinson-Fould; SM: Super Matrix.

Abstract (250 words)

Mitochondrial data has traditionally been used in reconstructing a variety of species phylogenies. The low rates of recombination and thorough characterization of mitochondrial data across vertebrate species makes it a particularly attractive phylogenetic marker. The relatively low number of fully sequenced mammal genomes and the lack of extensive sampling within Superorders have posed a serious problem for reaching agreement on the placement mammal species. The use of mitochondrial data sequences from large numbers of mammals could serve to circumvent the taxon-sampling deficit. Here we assess the suitability of mitochondrial data as a phylogenetic marker in mammal phylogenetics. MtDNA datasets of mammal origin have been filtered as follows: (i) we have sampled sparsely across the phylogenetic tree, (ii) we have constrained our sampling to genes with high taxon coverage, (iii) we have categorized rates across sites in a phylogeny independent manner and have removed fast evolving sites, and (iv), we have sampled from very shallow divergence times to reduce phylogenetic conflict. However, topologies obtained using these filters are not consistent with previous studies and are discordant across different genes. Individual mitochondrial genes, such as CO1 and CYTB currently used in the Barcode of life project, and indeed all mitochondrial genes analysed as a supermatrix resulted in poor resolution of the species phylogeny. Overall, our study highlights the limitations of mitochondrial data, not only for resolving deep divergences and but also shallow divergences in the mammal phylogeny.

Introduction

There are several differences between the nuclear and mitochondrial genome including but not restricted to: size of genome, mode of inheritance, levels and extent of recombination, number of introns and DNA repair mechanisms (Ballard and Whitlock 2004). Mitochondrial genes (mtGenes) undergo more mutations compared to nuclear genes (nucGenes) and are therefore more susceptible to saturation of base changes - a major challenge in phylogeny reconstruction (Brown et al. 1982). In contrast, the benefits of using mtGenes in phylogenetic studies are that mtGenes have very low rates of recombination (Awadalla et al. 1999; Hoarau et al. 2002; Ladoukakis and Zouros 2001; Lunt and Hyman 1997), mtGene order is relatively well conserved across vertebrates (Pereira 2000) and - specific to the focus of this paper - mtGene sequence data is available for over 1,000 mammals (UniProt 2012). The number of fully sequenced mammal nuclear genomes remains relatively low with 41 mammal genomes available in the Ensembl database (Flicek et al. 2014)out of ~5000 classified mammal species (Myers et al. 2014). Mammal phylogeneticists are therefore faced with severe restrictions on extensive taxon sampling within the Superorders. As mitochondrial sequences are readily available for so many taxa, the use of mitochondrial sequences could serve to ameliorate the taxon-sampling deficiency in nuclear sequences. Over the past number of years, studies have used both mitochondrial and nuclear genes to attempt to resolve the mammal phylogeny (Hallstrom and Janke 2008; Morgan et al. 2013; Nikolaev et al. 2007; Nishihara et al. 2006; Romiguier et al. 2013; Tobe et al. 2010) as well as morphological data (O'Leary et al. 2013) and rare genomic events (Nishihara et al. 2009).

MtGenes have previously been used to resolve deep phylogenetic relationships such as the placement of the Superorders in the mammal phylogeny (Gibson et al. 2005; Milinkovitch et al. 1993; Tobe et al. 2010), and also for more shallow relationships such as those amongst the Cetacea (Milinkovitch et al. 1993), the Caniforma (Arnason et al. 2007) and the Rodentia (Frye and Hedges 1995). The mitochondrial gene Cytochrome b (CYTB) was once the primary locus involved in phylogenetic studies (Irwin et al. 1991), but the Bar code of Life Consortium has adopted the mitochondrial gene Cytochrome c oxidase I (CO1) (Hebert et al. 2003) for the resolution of the eukaryote phylogeny. To date the most taxon rich phylogenetic study of mammals used the CYTB and CO1 genes and spanned 204 taxa (Tobe et al. 2010). This study revealed that while CYTB was a stronger candidate than CO1 for phylogeny reconstruction, neither gene could resolve the branching of the Superorders (Tobe et al. 2010). An analysis of the entire mitochondrial genome of 78 Eutherian taxa found strong support for the four Superorders of placental mammals (Kjer and Honeycutt 2007). However this study conflicted with nuclear DNA based studies as regards the

position of the Scandentia (Murphy et al. 2001a; Murphy et al. 2001b; Novacek 1992; Springer et al. 2004). Using nuclear DNA it had been resolved that Primate Orders are monophyletic (Murphy et al. 2001a; Murphy et al. 2001b), however using the entire mitochondrial genome a paraphyletic grouping of primates was retrieved, proposing a grouping of Dermoptera with anthropoid Primates to the exclusion of lineages such as tarsiers and prosimians (Kjer and Honeycutt 2007). These results are also incongruent with morphological studies for the position of these groups (O'Leary et al. 2013). These discrepancies signal that the application of mitochondrial data to the mammal phylogeny may be problematic. Here we apply a variety of assessments of data quality and signal to determine which (if any) mtGenes can be applied to mammal phylogenetics and at which phylogenetic depth.

A study of *Plethodon* salamanders showed that while incongruence between inferred mtDNA phylogenies was higher than inferred nucDNA phylogenies, the combined nuclear and mitochondrial data provided enough reliable phylogenetic signal that phylogenetic inconsistencies such as homoplasy and LBA present in the mitochondrial data were overcome (Fisher-Reid and Wiens 2011). And indeed this combined approach has been performed in the analysis of 66 Eutherian mammals using combined nuclear and mitochondrial data showed strong support for both Superorders and Orders (Murphy et al. 2001a). In summary, mitochondrial data appears to have performed well when combined with nuclear data in previous publications (Fisher-Reid and Wiens 2011; Murphy et al. 2001a).

Springer *et al*. (2001) carried out an investigation of the phylogenetic informativeness of mitochondrial versus nuclear gene sequences for deep-level mammal phylogeny reconstruction. They used the available data at the time, i.e. 32 taxa across 12 mitochondrial protein coding genes, together with a parsimony and minimum evolution approach (Springer et al. 2001). The conclusions were that concatenated nuclear genes were more effective at recovering benchmark clades compared with concatenated mtGenes alone (Springer et al. 2001). Since this initial study, there has been a surge in sequencing efforts and significant improvements to models and methods for phylogeny reconstruction of large datasets (Stamatakis 2006). Currently there are mitochondrial sequence data for over 1,000 placental mammals providing us with ample data to test if more mtDNA data improves the performance of this data type in reconstructing mammal phylogeny.

We sought to test the phylogenetic informativeness of each gene and ultimately identify the subset of mtGenes that provide the greatest phylogenetic information across a total of 455 placental mammal taxa. We assessed the phylogenetic congruence between individual mtGene phylogenies

and compared these to a phylogeny resolved from a dataset of concatenated mitochondrial genes. Phylogenetic conflict can arise from a number of features of the data and the method such as taxon sampling (Hedtke et al. 2006), lack of sufficient phylogenetic characters (Rosenberg and Kumar 2003) and saturation, resulting in homoplasy at deeper phylogenetic nodes (Caterino et al. 2001; Reed and Sperling 1999). We have assessed these phylogenetic conflicts within mitochondrial data by systematically reducing our dataset by taxa, by assessing the impact of gene coverage versus taxon sampling on phylogenetic informativeness, by removing rapidly evolving sites, and finally, by sampling sequence data at different depths on the known phylogenetic tree to assess where the phylogenetic signal starts to break down.

Materials and Methods

Gene and Taxon Sampling

Mitochondrion-encoded protein coding genes were downloaded for 1,556 taxa that spanned the four mammal Superorders (Euarchontoglires, Laurasiatheria, Xenarthra and Afrotheria) as well as non-mammal outgroup species (*Monodelphis domestica* and *Ornithorhynchus anatinus)* and Aves (*Gallus gallus*) from the UniProtKB database (UniProt 2012). Only taxa that were represented in at least two out of 13 mitochondrial genes (mtGenes) were used in this analysis, reducing the dataset to 455 taxa. For summary of data used in this analysis see Table 1 (full detail on individual taxon coverage is given in Supplementary Table 1).

Multiple Sequence Alignment

The 13 mtGene datasets were aligned using Muscle v3.7 (Edgar 2004) and quality was assessed using the norMD score (Thompson et al. 2001). All alignments (including the supermatrix (SM) dataset) had a norMD score > 0.6, indicating that the sequences in the MSA were well aligned and suitable for phylogenetic testing (Supplementary Table 2).

Model Choice and Phylogeny reconstruction

Model testing was performed using ModelGenerator v85 (Keane et al. 2006). RAxML (Stamatakis 2006) was employed for phylogeny reconstruction using the rapid bootstrapping algorithm (Stamatakis et al. 2007) where 1,000 bootstrap replicates were performed on each dataset using the best-fit model. A list of all models, log-likelihood (lnL) scores and phylogenetic trees are available in Supplementary Table 3.

Likelihood mapping tests

Likelihood mapping (LM) was performed on all datasets using TreePuzzle v5.2 (Schmidt et al. 2002). The mtMAM+4**Γ** model was not available in TreePuzzle v5.2 (Schmidt et al. 2002) so the next available model of best-fit defined through BIC analysis was chosen (usually mtREV+4Γ). LM analysis in TreePuzzle can only be performed on MSAs between four and 257 taxa. This is to avoid overflow of internal integer variables. Therefore, for datasets that exceeded this limit we randomly sampled 200 taxa 100 times from these datasets and have presented the mean scores from the LM analysis of these individual datasets. A full list of LM scores is given in Supplementary Table 4.

Removal of Saturated Sites

The rates of change of characters were categorized using TIGER (Cummins and McInerney 2011), a phylogeny independent method for classification of rates across sites. Twenty site categories were generated; where site category 1 represents characters associated with slowly evolving sites and site category 20 represents characters that are rapidly evolving. The sites that were associated with categories 20, 19 or 18 were removed in turn to generate "site-stripped" alignment. The site-stripped alignments generated from TIGER (Cummins and McInerney 2011) were assessed for phylogenetic signal using LM (Schmidt et al. 2002) and phylogenies were reconstructed using RAxML (Stamatakis 2006).

Calculate distance between topologies

To assess the levels of congruence between topologies, a majority rule (MR) consensus tree was generated using RAxML (Stamatakis 2006) and the Robinson-Foulds (RF) distance was calculated between two phylogenetic trees using the "rfdists" command in Clann (Creevey and McInerney 2005). The RF distance metric estimates the number of shared splits between the shared taxon set of two unrooted trees (Creevey and McInerney 2005). The numbers are reported as the ratio of the number of shared splits across the two trees, therefore a value of zero indicates that both trees share all splits while a number of one is given when the pair of trees shares no splits in common. Individual RF scores for all comparisons are detailed in Supplementary Table 5.

Results

Thirteen mitochondrial protein-coding genes were downloaded from the UniProtKB database (UniProt 2012). A total of 455 taxa had at least two sequences out of 13 mtGenes in the dataset. Taxa were sampled across 19 Placental Orders, Figure 1(A) (Meredith et al. 2011; Morgan et al. 2013). The 13 genes ranged in length from 71 amino acids (aa) to 626 aa and in taxon coverage from 94 to 281 taxa.

The phylogenetic conflict in these datasets was assessed using Likelihood Mapping (LM), which gives a prior indication of tree-likeness based on the distribution of likelihood vectors (Strimmer and von Haeseler 1997). The majority of signal is expected to fall within regions 1-3 – if there is strong phylogenetic signal and low levels of conflict, while signal falling within regions 4-6 is indicative of net-like relationships and signal in region 7 represents conflict, see Supplementary Table 3. Strimmer and Haeseler (1997) simulated datasets of different lengths and showed that a cumulative percentage of 8.5% from regions 4 through 7 produced a bifurcating tree for an alignment of 200 base pairs (bp). We therefore defined a cut off of <10% phylogenetic conflict for all our datasets as these alignments should produce reasonably well supported bifurcating trees. Datasets with a high proportion of phylogenetic conflict (> 10%) were expected to produce less well-resolved nodes due to the remaining data contributing to tree-likeness (Table 1) (Strimmer and von Haeseler 1997). In total, 11/13 mtGenes had a cumulative score across regions 4 through 7 in the LM analysis of > 10% indicating a level of phylogenetic conflict above our acceptance level. The two genes that satisfied our criteria of < 10% conflict were ND4 (9.7% conflict) and ND5 (8.1% conflict). In addition to analysing each mtGene individually, the mtGenes were concatenated to form a Supermatrix (SM) consisting of 3,906 aa and 455 taxa. Phylogeny reconstruction was carried out in a ML framework using RAxML (Stamatakis 2006).

The resultant phylogenies from both the individual gene analyses and SM dataset contained large numbers of weak and un-supported nodes. Congruence between majority rule consensus topologies was assessed using Robinson-Foulds (RF) distance as implemented in the Clann software (Creevey and McInerney 2005) (Supplementary Table 5). The results showed that the topology obtained from the ND5 gene was the closest to the topology obtained using the SM dataset, with a RF distance of 0.1301. The two genes used in the Barcode of Life project (Borisenko et al. 2008; Hebert et al. 2003), CYTB and CO1, manifested RF distances of 0.2140 and 0.2609 respectively when compared to the topology obtained from the SM dataset and the CYTB and CO1 gene trees had an RF distance of 0.2021 to one another. While it is widely accepted that the placental mammals are

grouped into four Superorders (Meredith et al. 2011; Morgan et al. 2013), Figure 1A, here we observed that none of the datasets generated from mtGenes, i.e. neither individual gene datasets nor the SM dataset, were able to resolve these four Superorders (Figure 1B). MtDNA accumulates mutations more rapidly than nuclear data, and therefore is more likely to have both saturation and homoplasy (Brown et al. 1982; Rubinoff and Holland 2005), both of which contribute to phylogenetic conflict. This has resulted in inconsistencies between phylogenies generated from nuclear and mitochondrial data (Caterino et al. 2001; Reed and Sperling 1999; Rokas and Carroll 2008). In an effort to reduce phylogenetic conflict, increase node support and improve upon congruence between mtGene topologies a number of issues were addressed. First, the phylogenetic conflict was assessed to see if it decreased with a reduction in taxon number. Then we assessed whether phylogenetic signal is stronger when gene coverage across taxa is higher. The impact of the removal of saturated sites was assessed, as was the impact of node depth on phylogenetic signal. To answer each of these questions the data were subjected to a series of treatments and the outcome in each case is detailed below.

Phylogenetic conflict does not decrease with a reduction in the number of taxa

It has been debated whether more sequence data or more thorough sampling improves phylogeny reconstruction (Hedtke et al. 2006; Hillis et al. 2003; Pollock et al. 2002; Rosenberg and Kumar 2001, 2003). To test the impact of reduced taxon sampling on phylogenetic signal, a subset of taxa were sampled (between nine and 13 species) for each of the mtGenes. In each case a representative from each placental mammal Superorder was retained in the dataset. The reduced taxon datasets were re-tested for phylogenetic conflict using LM (Schmidt et al. 2002). From this analysis, it was observed that there was no individual gene that when removed from the dataset showed a significant reduction in phylogenetic conflict (Supplementary Table 4). More specifically, conflict increased in 12 out of 13 mtGenes, the only exception was CO1 that manifested a small reduction from 18.5% to 17.3% conflict. The SM dataset, with reduced taxon sampling, showed the lowest level of conflict of all the datasets with a conflict score of 3.4%. Phylogenetic reconstruction of the treated SM dataset was expected to resolve four placental Superorders with platypus positioned as outgroup (van Rheede et al. 2006). However, there were only low levels of support for the four placental Superorders and there was 97% bootstrap support for a relationship joining Opossum and Platypus as sister taxa to the exclusion of all other mammals. Regardless of the strategy of restricted sampling from the Superorders, the data were still unable to provide support for the placement of four placental mammal Superorders. Therefore the reduction in taxa sampled from the mtGene data did not reduce phylogenetic conflict or improve phylogenetic resolution. The phylogenetic inconsistencies may have resulted from missing data.

Phylogenetic signal is stronger when gene coverage across taxa is higher

MtGenes have been sequenced to varying extents across placental mammals, and only 25 taxa have been sequenced for all 13 mtGenes. Congruence between phylogenies indicates how much error is contained in each phylogeny (Pisani et al. 2007). Missing sequence data is problematic in phylogeny reconstruction (Kearney 2002; Lemmon et al. 2009), however if enough phylogenetically informative characters are available then missing sequence data does not impact accurate phylogeny reconstruction (Philippe et al. 2004; Wiens 2003). Consequently, our next approach was to determine the impact of increasing gene coverage across the data. We increased the gene coverage gradually from two to 13 genes, and at each step generated a dataset (consequently the number of taxa decreased at each step). The SM dataset and the individual mtGene datasets were treated in this way.

LM (Schmidt et al. 2002) was employed to test the change in phylogenetic signal as gene coverage was increased (Figure 2). Phylogenetic conflict remained extremely high in ATP6, ATP8, CO1, CO2, CO3, ND3, ND4L and ND6 across all datasets regardless of gene coverage. ND1 showed variable phylogenetic conflict (12.2 - 14.9%) across the different levels of gene coverage but failed to reach our pre-defined cut-off value of < 10% conflict. CYTB, ND2, and ND5 showed < 10% phylogenetic conflict with the highest gene coverage and lowest taxon coverage conditions (Figure 2). ND5 maintained reasonably low phylogenetic conflict across all gene coverage situations (5.8-8.6% conflict). The RF distance was calculated between ND5 gene topologies and topologies from other mtGene datasets and the SM dataset to assess if congruence between gene trees improved at any coverage point. It was expected that if the datasets had more taxa in common, then the topological distance between gene topologies would be smaller. The RF distances showed that when gene coverage was at its lowest (*i.e.* just two mtGenes) then the ND5 gene had the closest RF distance between seven other mtGenes (CO1, CO2, CO3, ND1, ND4, ND6) and the SM topologies (Supplementary Table 5). Therefore, maximising the gene coverage across genes to improve congruence in these data does not have the expected effect. Only the *Glires*, *Carnivora* and *Cetartiodactlya* are represented in the 13 mtGene set, and so resolution of other clades is not possible with current data.

Upon examination of the SM dataset, there was a notable trend towards a decrease in phylogenetic conflict, from 7.2% to 1.1%, as gene coverage increased and taxa number decreased (Figure 2). This is unsurprising given longer sequences (e.g. concatenated alignments) increase the number of usable characters and that has been shown to overcome the phylogenetic inconsistencies of individual gene data (Gadagkar et al. 2005).

To test the quality of the phylogenetic signal, ML trees were drawn from the SM dataset across all gene coverage levels (Supplementary Table 3). The topologies do not reflect trends in LM tests, as improvement in node support is not observed with decrease in phylogenetic conflict. When gene coverage is between two and four genes, there are multiple collapsed nodes (branch support < 50%), which is indicative of large proportions of phylogenetic conflict (Supplementary Table 3). Four clearly defined Superorders were observed when gene coverage was exactly 4 and also when it was between six and nine genes, with a range of 109 to 284 taxa (Supplementary Table 3). The topological distance between phylogenies for each mtGene dataset and the SM dataset were calculated using RF distances at each level of gene coverage. It was found that there was no exact agreement between topologies (RF > 0.00) from individual mtGenes and the SM datasets for the same gene coverage. While an increase in gene coverage and a decrease in missing data provided sufficient signal to resolve the four Superorders, strong node support for intra-ordinal nodes was not achieved using these data.

Removal of saturated sites does not reduce the conflict in mitochondrial data

Mitochondrial datasets tend to have more saturation compared to nuclear datasets (Brown et al. 1982). In an effort to identify and remove rapidly evolving or saturated sites from the data, sites were categorised based on their rates of evolution using the phylogeny independent method TIGER (Cummins and McInerney 2011) and the ML phylogeny-dependent method implemented in TreePuzzle (Schmidt et al. 2002). LM was performed at each stepwise reduction in alignment length, and changes in the level of phylogenetic conflict were assessed (Supplementary Table 4).

The TIGER (Cummins and McInerney 2011) method showed that when the fastest site category was removed (site category 20), a slight reduction in phylogenetic conflict was observed for ATP8 (36.6% to 35.4%) and ND5 (8.1% to 8.0%), but there was no change in phylogenetic conflict observed in the ATP6 gene (17.8%) for the same manipulation. The removal of site category 20 resulted in an increase in phylogenetic conflict for the remaining 10 mtGenes, suggesting that removal of site category 20 could be removing necessary phylogenetic signal. Subsequent removals of site categories, e.g. site categories [20 and 19] and site categories [20, 19 and 18], resulted in an increase in the phylogenetic conflict in all 13 mtGenes. Removal of site category 20 from the SM dataset reduced the concatenated alignment from 4329 aa to 882 aa. Unfortunately, this reduction in sequence length left too few of overlapping characters per taxa for phylogeny reconstruction to be carried out.

Phylogenies were generated at each step for the individual mtGene datasets. However, as the fast evolving site categories were stepwise removed, this resulted in a reduction in the number of bifurcating nodes in the resultant phylogeny. A profile of the frequency of amino acids occurring under each site category estimated is provided in Figure 3. For each of the mtGenes, a large proportion of site categories were categorised as highly conserved (categories 1-3) or rapidly evolving (categories 18-20) with an average of 39.15% of sites (categories 4-17) sites remaining for phylogeny reconstruction. Phylogenies and LM results from the TIGER (Cummins and McInerney 2011) analyses have been provided in Supplementary Table 3 and 4 respectively.

Phylogenetic signal does not improve at more shallow divergence times

Previous studies have shown that high levels of homoplasy are observed when sampling from deep nodes using mitochondrial data (Caterino et al. 2001; Reed and Sperling 1999). The aim of this part of the analysis was to understand precisely at which depth the phylogenetic signal starts to degrade when using mtGenes. Groups of taxa were selected at different depths on the known species phylogeny (Meredith et al. 2011; Morgan et al. 2013) (Figure 4(A)). The closest available species were chosen as outgroups for each subset of data. Phylogenetic conflict was estimated from each dataset using LM (Schmidt et al. 2002) and all topologies were generated using RAxML (Stamatakis 2006). The levels of phylogenetic conflict varied over the 13 mtGenes depending on node depth. A summary of datasets that passed the < 10% phylogenetic conflict cut-off are shown in Figure 4 (detailed LM results are available in Supplementary Table 4).

A decrease in phylogenetic conflict was observed at shallower phylogenetic depths for ATP6, COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND5 and the SM, however, no improvement in tree-likeness (phylogenetic conflict > 10% at all nodes) was observed for mtGenes ATP8, ND4L and ND6. While we observed less phylogenetic conflict when nodes were sampled from shallower depths on the known species tree for some of the data, phylogenetic conflict did not decrease uniformly from deep to shallow nodes, e.g. the phylogenetic conflict for ND4 was as follows: *Eutherian* node (9.7% conflict), *Boreoeutheria* node (8.2% conflict), *Euarchontoglires* node (8.9% conflict) and Primates (5.8% conflict). Sampling the mtGene ND4 at the node defining the *Eutherian* ancestor and comparing the resultant topology with those generated from data sampled at shallower nodes, the distance between the trees varies as follows: *Boreoeutheria* node (RF distance = 0.0176), *Euarchontoglires* node (RF distance = 0.0294) and Primates (RF distance = 0.0405). While small improvements in tree-likeliness are observed from sampling taxa at shallower nodes this methodology does not produce a consistent results as congruence between sub sampled data is not observed.

The most successful shallow nodes for producing tree-likeness were Primates, Cetartiodactyla, Perissodactyla, Carnivora and Afrotheria. We produced a SM using the best performing mtGenes, CYTB, ND1, ND2 and ND5 which resulted in a LM conflict score of 7.27% but a phylogenetic tree that showed the Primates as paraphyletic thus not fully resolving the tested taxa into their associated Superorders, *Euarchontoglires* and *Laurasiatheria*.

Overall there are considerable levels of variation in the topological findings and there is more discordance between the phylogenies from the mtGenes and the SuperMatrix datasets than there is topological congruence.

Discussion

Previous phylogenetic studies of mitochondrial data show that homoplasy is not as prevalent at shallower nodes (Caterino et al. 2001; Reed and Sperling 1999). Here we find phylogenetic conflict in mtGene data at both deep and shallow nodes calling into question the use of mtDNA in phylogenetic studies of mammals at all levels. According to our results, none of the mtGenes were determined to be good candidates for phylogenetic reconstruction. This includes CYTB and CO1 currently used in the bar code of life project (Borisenko et al. 2008; Hebert et al. 2003) and previously held as a valid approach for resolving the phylogeny of rodent species such as the *Praomyini* (Nicolas et al. 2012). While there are a number of individual cases where using CYTB and CO1 as phylogenetic markers have been successful (Nicolas et al. 2012), preference has been awarded to CO1 as a phylogenetic marker over other mitochondrial genes (Luo et al. 2011) and studies have pointed out the usefulness of these data to pinpoint misidentified species (Shen et al. 2013). In this study we have observed levels of incongruence that call into question the application of any mitochondrial gene for taxonomic identification.

The root of the placental mammal tree has been widely contested of late (Morgan et al. 2013; Romiguier et al. 2013; Teeling and Hedges 2013) and so it was unsurprising to see variations in the position of the *Xenarthra* and the *Afrotheria* at the base of the placental tree. The four Superorders of placental mammals are observed by multiple independent studies using nuclear data (Hallstrom and Janke 2008; Meredith et al. 2011; Murphy et al. 2001b), rare genomic change (Murphy et al. 2007), nuclear and mitochondrial data combined (Murphy et al. 2001a) and a study that used the entire mitochondria genome on 78 taxa (Kjer and Honeycutt 2007). The SM dataset applied here displayed less phylogenetic conflict than the individual gene datasets, but the four well-defined

Superorders were not supported. While longer alignments have been shown to overcome phylogenetic inconsistencies of smaller datasets our results suggest that this is not always the case (Gadagkar et al. 2005; Gee 2003). Likewise, previous large-scale phylogenomic studies have found phylogenetic inconsistencies regardless of implementation of large Supermatrix (SM) style datasets (Dunn et al. 2008; Philippe et al. 2009; Schierwater et al. 2009). Phylogenomic studies of mammals have attributed this inconsistency to introgression or gene flow as a result of hybridization (Hallstrom and Janke 2008). The observations from Hallstrom and Janke (2008) were based on nuclear data. Introgression in mtGenes has been identified within species of mammals such as the *Canis* genus (Hailer and Leonard 2008) and full mitochondrial genome replacement has been shown within the Chiroptera Order (Berthier et al. 2006). It is possible that these evolutionary phenomena acting on mtGenes are negatively impacting the accurate resolution of the phylogenetic history of mammals.

There is an array of opinions on the impact of missing data on phylogeny reconstruction (Kearney 2002; Lemmon et al. 2009; Philippe et al. 2004; Wiens 2003). In this study small improvements were observed when increasing gene coverage across the SM dataset with regards to the placement of the Superorders but phylogenetic conflict was still observed at shallower nodes. Removal of fast evolving sites from mtGene sequence data did not reduce the phylogenetic conflict nor did it improve overall resolution of the phylogeny. Incongruence between mtGene phylogenies is an indicator of the level of error between two trees (Pisani et al. 2007) and as high levels of incongruence have been observed throughout this study (regardless of whether the data was treated or not), it does not increase our confidence in the application of mtGenes as a phylogenetic marker in mammal studies.

While congruence in phylogenies generated from mtGene data is important, so too is congruence between different data types such as nuclear sequences, morphological data and rare genomic elements (Branger et al. 2011; Campbell et al. 2011; Pisani et al. 2007; Rota-Stabelli et al. 2011). Once again, the mtGene data was unable to generate topologies that agreed with previous studies of different data types (Meredith et al. 2011; Morgan et al. 2013; Murphy et al. 2007; Shoshani et al. 1996), and differed in the resolution of the four Superorders and inter-ordinal placements.

Previously, caution has been issued against phylogenetic reconstruction using exclusively mitochondrial data (Rubinoff and Holland 2005; Shaw 2002), and others have supported the use of a single mtGene (CO1) for taxonomic placement (Luo et al. 2011). Here we demonstrate that mtGenes are not suitable for resolving the mammal phylogeny. While improvements are observed

upon treating the mtGene data using various partitioning techniques, the resultant topologies are incongruent with the well-known Superorder groupings (Figure 1(A)). Using individual genes is not recommended for further topological evaluations of the placental mammals; this includes those genes used in the bar code of life project (CO1 and CYTB).

Supplementary Tables

Supplementary Table 1: Taxon coverage across mtGenes.

The species name is given along with whether or not it is represented in each of the 13 mtGenes. The final column shows the total number of times the species is represented across all mtGenes.

Supplementary Table 2: MSA for untreated mtGene and SM datasets.

All alignments used in this study are supplied in this file.

Supplementary Table 3: Phylogenetic trees obtained for all Datasets.

The dataset is listed along with the phylogenetic tree and its associated lnL score. The Γ parameter is denoted as +G throughout.

Supplementary Table 4: Summary of Likelihood Mapping for all Datasets

For each dataset, the number of taxa and amino acids are given along with the scores for regions 1-7 from LM analysis. The phylogenetic conflict score is the sum of values from regions 4-7 and this is given in the final column.

Supplementary Table 5: Robinson-Fold distances between topologies.

Figure Legends

Figure 1 Phylogeny inferred from nuclear and mitochondrial data.

The phylogeny obtained using (A) nuclear data and (B) mitochondrial data is shown. The accepted Superorders of the placental mammals are colour coded according to the following scheme: *Euarchontoglires* = red, *Laurasiatheria* = blue, *Afrotheria* = green and *Xenarthra* = purple.

Figure 2 The impact of Gene Coverage versus Taxon Sampling on phylogenetic Signal as measured by percentage of phylogenetic conflict

The rows represent datasets generated from the individual mtGenes and the SM dataset. The columns represent gene coverage across taxa (from 13 to 2 genes) for each dataset and the numbers in each cell represent the number of taxa in a given dataset. The percentage of phylogenetic conflict is colour coded as shown in (B) from acceptable levels (< 10% conflict) represented by pale yellow and green, to unacceptable levels (≥ 10% conflict) represented by orange and red.

Figure 3: A profile of the distribution of site-categories across the datasets.
The frequency of amino acids (y-axis) that are estimated to be evolving at a rate corresponding to a given "site-category" depicted on the x-axis as site categories (or Bins) 1 – 20 (*i.e.* from slowest to fastest evolving). The results of this site-categorization are shown for each of the untreated mtGenes (A –M) and for the SM dataset (N).

Figure 4: Assessing phylogenetic conflict in datasets sampled at different depths on the known placental mammal phylogeny.

Panel (A) shows nodes circled with grey that were tested in the analysis and numbers within these circles represent how many mtGenes support the tree-likeliness of that node. A summary table of which genes support each node is provided to the left of panel (A). Each phylogenetic tree (B-I) represents the analysis of a mtGene as labelled, and (J) represents the Supermatrix (SM) dataset. The representative taxa used in each dataset (A-J) are identical. The bolded lines represent either Superorders or Orders where the phylogenetic conflict was < 10%. [1]Caniforma and [2]Cetacea denotes where these Orders within their Superorders also passed cut-off criteria of < 10% phylogenetic conflict.

Table Legends

Table 1: Details of untreated mitochondrial data, model choice and Likelihood Mapping results.

The total number of taxa, and the sequence lengths are given for each untreated dataset along with their associated models of evolution and lnL values for the phylogenies generated through RAxML (Stamatakis 2006). The column on the left is the phylogenetic conflict score, i.e. the cumulative score from regions 4 through 7 inclusive from the LM analysis.

Author contributions

CCM carried out all data assembly, phylogenetic and statistical analyses. CJC and MJO'C contributed to methodology, phylogenetic tests and statistical analyses. CCM and MJO'C conceived of the study, its design and coordination. CJC, CCM and MJO'C all participated in drafting the manuscript.

Conflict of Interest

Authors declare no conflict of interest.

Acknowledgements

We would like to thank the Irish Research Council for Science, Engineering and Technology for the Embark Initiative Postgraduate Scholarship to CCM: RS2000172 and Science Foundation Ireland (SFI) for funding to Dr Mary J. O'Connell (EOB: 2673). We would like to thank the SFI/Higher Education authority (HEA) Irish Centre for High-End Computing (ICHEC: dclif023b) and SCI-SYM DCU for processor time. We would like to thank Paul Kilroy-Glynn for initial discussion, Dr Davide Pisani and Prof James McInerney for their helpful comments and the Orla Benson travel award (DCU) for funding.

Bibliography

Arnason U, Gullberg A, Janke A, Kullberg M (2007) Mitogenomic analyses of caniform relationships. Mol Phylogenet Evol 45, 863-874
Awadalla P, Eyre-Walker A, Smith JM (1999) Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286, 2524-2525
Ballard JW, Whitlock MC (2004) The incomplete natural history of mitochondria. Mol Ecol 13, 729-744
Berthier P, Excoffier L, Ruedi M (2006) Recurrent replacement of mtDNA and cryptic hybridization between two sibling bat species Myotis myotis and Myotis blythii. Proc Biol Sci 273, 3101-3109
Borisenko AV, Lim BK, Ivanova NV, Hanner RH, Hebert PD (2008) DNA barcoding in surveys of small mammal communities: a field study in Suriname. Mol Ecol Resour 8, 471-479
Branger B, Gillard P, Monrigal C, Thelu S, Robidas E, Viot S, Descamps P, Philippe HJ, Sentilhes L, Winer N (2011) [Lessons and impact of two audits on postpartum hemorrhages in 24 maternity hospitals of the network "Securite Naissance - Naitre Ensemble" in "Pays-de-la-Loire" area]. J Gynecol Obstet Biol Reprod (Paris) 40, 657-667
Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18, 225-239
Campbell LI, Rota-Stabelli O, Edgecombe GD, Marchioro T, Longhorn SJ, Telford MJ, Philippe H, Rebecchi L, Peterson KJ, Pisani D (2011) MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. Proc Natl Acad Sci U S A 108, 15920-15924
Caterino MS, Reed RD, Kuo MM, Sperling FA (2001) A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera:Papilionidae). Syst Biol 50, 106-127
Creevey CJ, McInerney JO (2005) Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21, 390-392
Cummins CA, McInerney JO (2011) A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst Biol 60, 833-844
Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SH, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745-749
Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792-1797
Fisher-Reid MC, Wiens JJ (2011) What are the consequences of combining nuclear and mitochondrial data for phylogenetic analysis? Lessons from Plethodon salamanders and 13 other vertebrate clades. BMC Evol Biol 11, 300
Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kahari AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ruffier M, Sheppard D, Taylor K, Thormann A, Trevanion SJ, Vullo A, Wilder SP, Wilson M, Zadissa A, Aken BL, Birney E, Cunningham F, Harrow J, Herrero J, Hubbard TJ, Kinsella R, Muffato M, Parker A, Spudich G, Yates A, Zerbino DR, Searle SM (2014) Ensembl 2014. Nucleic Acids Res 42, D749-755
Frye MS, Hedges SB (1995) Monophyly of the Order Rodentia Inferred from Mitochondrial-DNA Sequences of the Genes for 12s Ribosomal-Rna, 16s Ribosomal-Rna, and Transfer-Rna-Valine. Molecular Biology and Evolution 12, 168-176
Gadagkar SR, Rosenberg MS, Kumar S (2005) Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J Exp Zool B Mol Dev Evol 304, 64-74

Gee H (2003) Evolution: ending incongruence. Nature 425, 782
Gibson A, Gowri-Shankar V, Higgs PG, Rattray M (2005) A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol Biol Evol 22, 251-264
Hailer F, Leonard JA (2008) Hybridization among three native North American Canis species in a region of natural sympatry. Plos One 3, e3333
Hallstrom BM, Janke A (2008) Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations. BMC Evol Biol 8, 162
Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc Biol Sci 270, 313-321
Hedtke SM, Townsend TM, Hillis DM (2006) Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Systematic Biol 55, 522-529
Hillis DM, Pollock DD, McGuire JA, Zwickl DJ (2003) Is sparse taxon sampling a problem for phylogenetic inference? Syst Biol 52, 124-126
Hoarau G, Holla S, Lescasse R, Stam WT, Olsen JL (2002) Heteroplasmy and evidence for recombination in the mitochondrial control region of the flatfish Platichthys flesus. Mol Biol Evol 19, 2261-2264
Irwin DM, Kocher TD, Wilson AC (1991) Evolution of the cytochrome b gene of mammals. J Mol Evol 32, 128-144
Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McLnerney JO (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol 6, 29
Kearney M (2002) Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Syst Biol 51, 369-381
Kjer KM, Honeycutt RL (2007) Site specific rates of mitochondrial genomes and the phylogeny of eutheria. BMC Evol Biol 7, 8
Ladoukakis ED, Zouros E (2001) Recombination in animal mitochondrial DNA: evidence from published sequences. Mol Biol Evol 18, 2127-2131
Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst Biol 58, 130-145
Lunt DH, Hyman BC (1997) Animal mitochondrial DNA recombination. Nature 387, 247
Luo A, Zhang A, Ho SY, Xu W, Zhang Y, Shi W, Cameron SL, Zhu C (2011) Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals. BMC genomics 12, 84
Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simao TL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ (2011) Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334, 521-524
Milinkovitch MC, Orti G, Meyer A (1993) Revised phylogeny of whales suggested by mitochondrial ribosomal DNA sequences. Nature 361, 346-348
Morgan CC, Foster PG, Webb AE, Pisani D, McInerney JO, O'Connell MJ (2013) Heterogeneous models place the root of the placental mammal phylogeny. Mol Biol Evol 30, 2145-2156
Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ (2001a) Molecular phylogenetics and the origins of placental mammals. Nature 409, 614-618
Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS (2001b) Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294, 2348-2351
Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W (2007) Using genomic data to unravel the root of the placental mammal phylogeny. Genome Research 17, 413-421

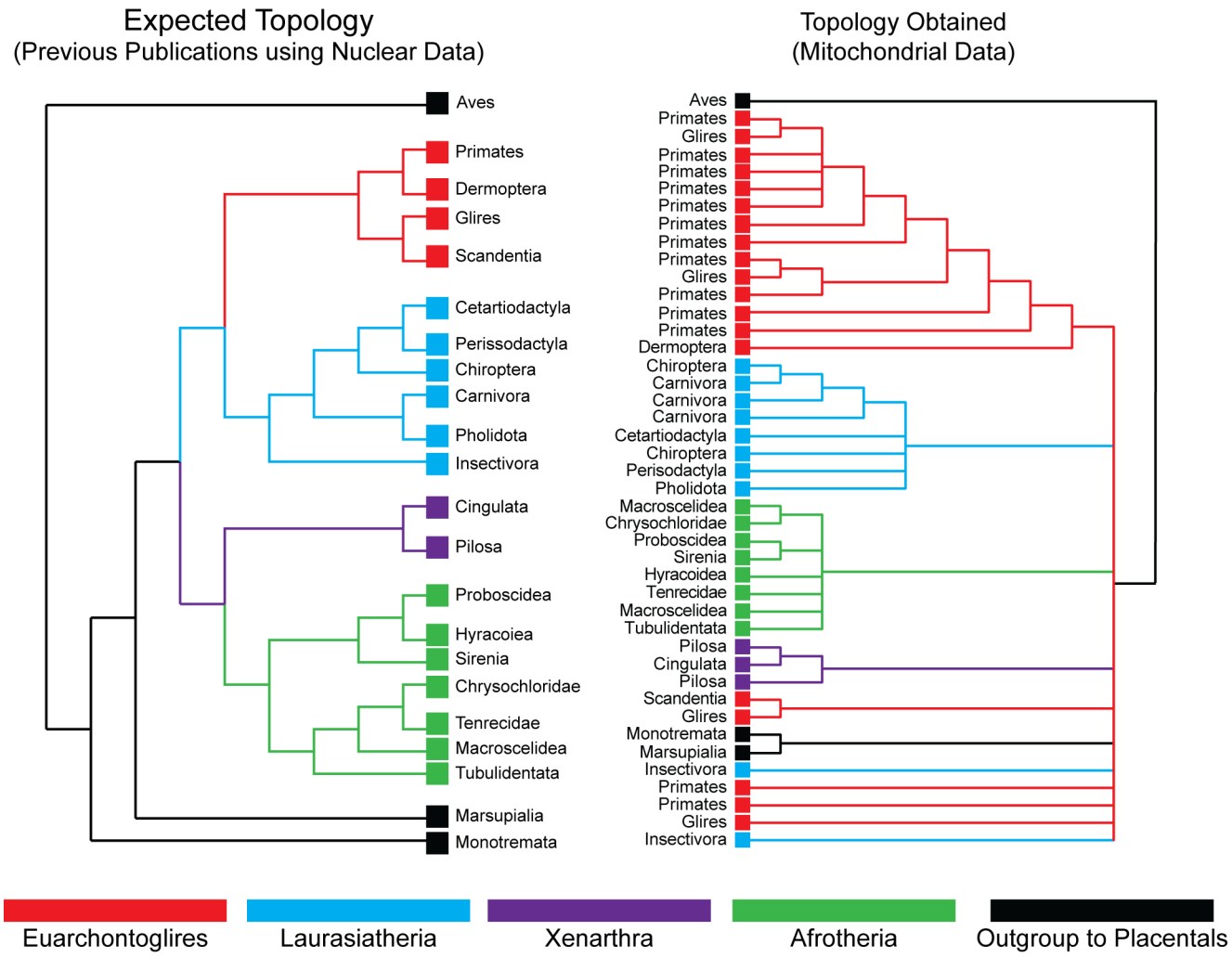
Myers P, Espinosa R, Parr CS, Jones T, Hammond GS, Dewey TA (2014) The Animal Diversity Web (online). Accessed at http://animaldiversity.org.
Nicolas V, Schaeffer B, Missoup AD, Kennis J, Colyn M, Denys C, Tatard C, Cruaud C, Laredo C (2012) Assessment of three mitochondrial genes (16S, Cytb, CO1) for identifying species in the Praomyini tribe (Rodentia: Muridae). PLoS One 7, e36586
Nikolaev S, Montoya-Burgos JI, Margulies EH, Rougemont J, Nyffeler B, Antonarakis SE (2007) Early history of mammals is elucidated with the ENCODE multiple species sequencing data. Plos Genetics 3, e2
Nishihara H, Hasegawa M, Okada N (2006) Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. Proc Natl Acad Sci U S A 103, 9929-9934
Nishihara H, Maruyama S, Okada N (2009) Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. Proc Natl Acad Sci U S A 106, 5235-5240
Novacek MJ (1992) Mammalian phylogeny: shaking the tree. Nature 356, 121-125
O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo ZX, Meng J, Ni X, Novacek MJ, Perini FA, Randall ZS, Rougier GW, Sargis EJ, Silcox MT, Simmons NB, Spaulding M, Velazco PM, Weksler M, Wible JR, Cirranello AL (2013) The placental mammal ancestor and the post-K-Pg radiation of placentals. Science 339, 662-667
Pereira SL (2000) Mitochondrial genome organization and vertebrate phylogenetics. Genetics and Moleculat Biology 23, 754-752
Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Queinnec E, Da Silva C, Wincker P, Le Guyader H, Leys S, Jackson DJ, Schreiber F, Erpenbeck D, Morgenstern B, Worheide G, Manuel M (2009) Phylogenomics revives traditional views on deep animal relationships. Curr Biol 19, 706-712
Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, Casane D (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21, 1740-1752
Pisani D, Benton MJ, Wilkinson M (2007) Congruence of morphological and molecular phylogenies. Acta Biotheor 55, 269-281
Pollock DD, Zwickl DJ, McGuire JA, Hillis DM (2002) Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol 51, 664-671
Reed RD, Sperling FA (1999) Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus Papilio. Mol Biol Evol 16, 286-297
Rokas A, Carroll SB (2008) Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol 25, 1943-1953
Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJ (2013) Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol Biol Evol 30, 2134-2144
Rosenberg MS, Kumar S (2001) Incomplete taxon sampling is not a problem for phylogenetic inference. Proc Natl Acad Sci U S A 98, 10751-10756
Rosenberg MS, Kumar S (2003) Taxon sampling, bioinformatics, and phylogenomics. Syst Biol 52, 119-124
Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ (2011) A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc Biol Sci 278, 298-306
Rubinoff D, Holland BS (2005) Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. Syst Biol 54, 952-961
Schierwater B, Eitel M, Jakob W, Osigus HJ, Hadrys H, Dellaporta SL, Kolokotronis SO, Desalle R (2009) Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. Plos Biology 7, e20
Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18, 502-504

Shaw KL (2002) Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. Proc Natl Acad Sci U S A 99, 16122-16127
Shen YY, Chen X, Murphy RW (2013) Assessing DNA barcoding as a tool for species identification and data quality control. PLoS One 8, e57125
Shoshani J, Groves CP, Simons EL, Gunnell GF (1996) Primate phylogeny: morphological vs. molecular results. Mol Phylogenet Evol 5, 102-154
Springer MS, DeBry RW, Douady C, Amrine HM, Madsen O, de Jong WW, Stanhope MJ (2001) Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. Mol Biol Evol 18, 132-143
Springer MS, Stanhope MJ, Madsen O, de Jong WW (2004) Molecules consolidate the placental mammal tree. Trends Ecol Evol 19, 430-438
Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688-2690
Stamatakis A, Auch AF, Meier-Kolthoff J, Goker M (2007) AxPcoords & parallel AxParafit: statistical co-phylogenetic analyses on thousands of taxa. Bmc Bioinformatics 8, 405
Strimmer K, von Haeseler A (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci U S A 94, 6815-6819
Teeling EC, Hedges SB (2013) Making the impossible possible: rooting the tree of placental mammals. Mol Biol Evol 30, 1999-2000
Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O (2001) Towards a reliable objective function for multiple sequence alignments. J Mol Biol 314, 937-951
Tobe SS, Kitchener AC, Linacre AM (2010) Reconstructing mammalian phylogenies: a detailed comparison of the cytochrome B and cytochrome oxidase subunit I mitochondrial genes. PLoS One 5, e14156
UniProt (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40, D71-75
van Rheede T, Bastiaans T, Boone DN, Hedges SB, de Jong WW, Madsen O (2006) The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and Therians. Mol Biol Evol 23, 587-597
Wiens JJ (2003) Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol 52, 528-538

Table 1: Details of untreated mitochondrial data and model choice.

mtGene Name	Taxa #	MSA Length (aa)	Model of Evolution	-lnL	Conflict [4-7]
ATP6	253	228	MtMam+I+4Γ	-13653.63	17.80
ATP8	281	71	MtMam+I+4Γ	-9145.23	36.60
CO1	187	518	MtMam+I+4Γ	-7530.62	18.50
CO2	217	237	MtMam+4Γ	-6430.97	19.50
CO3	189	269	MtMam+I+4Γ	-7175.67	14.30
CYTB	267	383	MtMam+I+4Γ	-23093.23	12.20
ND1	129	326	MtMam+4Γ	-12503.65	14.00
ND2	152	350	MtMam+4Γ	-27716.40	12.40
ND3	141	119	MtMam+4Γ	-5619.87	25.50
ND4	163	486	MtMam+4Γ	-25191.86	9.70
ND4L	246	98	MtMam+4Γ	-7264.63	25.20
ND5	149	626	MtMam+I+4Γ+F	-41499.46	8.10
ND6	94	200	JTT+4Γ+F	-10035.93	18.40
SM	455	3906	MTMam+G+F	-204073.11	12.72

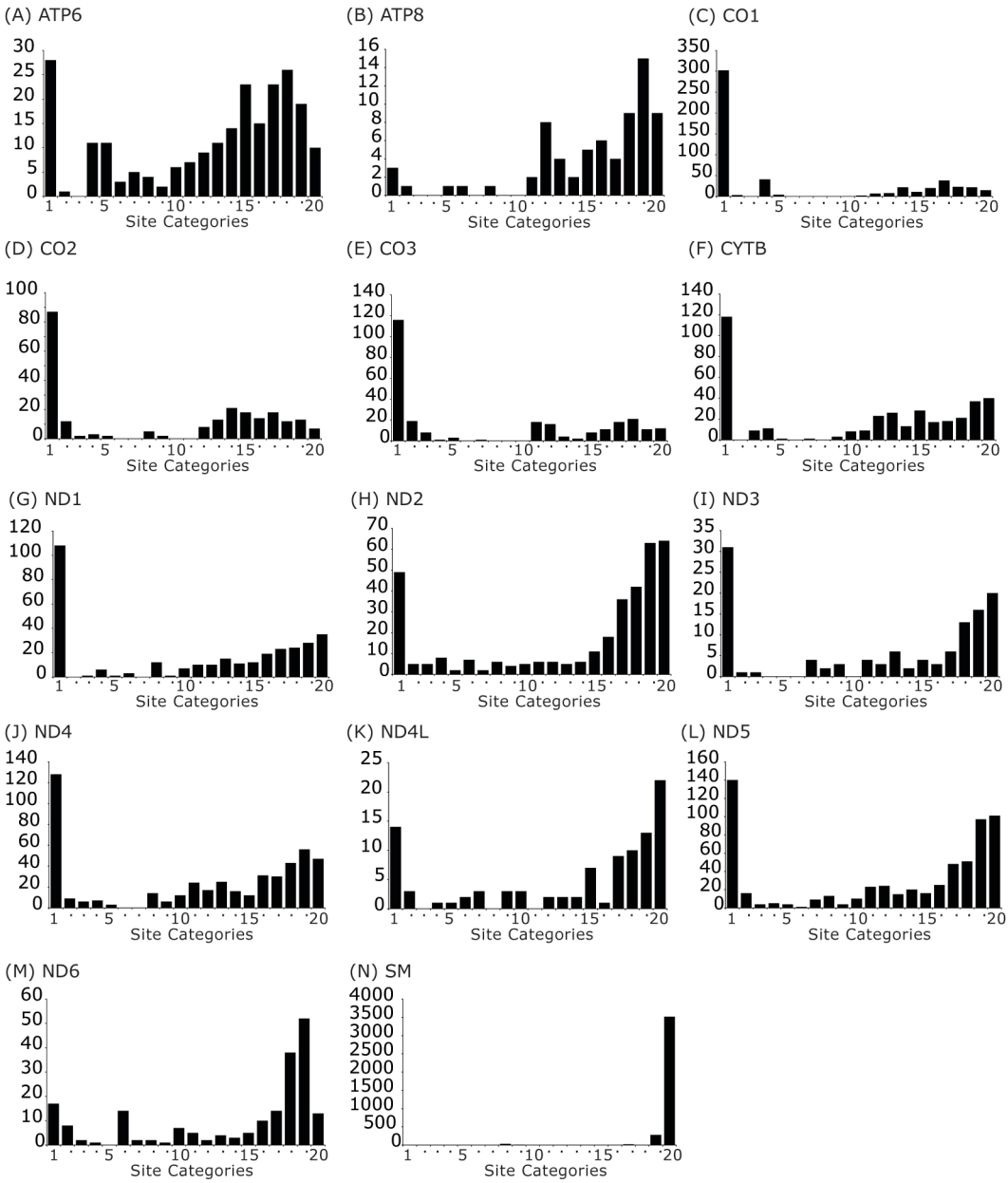
The total number of taxa, sequence length are given for each dataset along with their associated models of evolution and lnL values for phylogeny generated through RAxML (Stamatakis 2006).


Expected Topology
(Previous Publications using Nuclear Data)
Aves
Primates
Dermoptera
Glires
Scandentia
Cetartiodactyla
Perissodactyla
Chiroptera
Carnivora
Pholidota
Insectivora
Cingulata
Pilosa
Proboscidea
Hyracoiea
Sirenia
Chrysochloridae
Tenrecidae
Macroscelidea
Tubulidentata
Marsupialia
Monotremata
Topology Obtained
(Mitochondrial Data)
Aves
Primates
Glires
Primates
Primates
Primates
Primates
Primates
Primates
Primates
Glires
Primates
Primates
Primates
Dermoptera
Chiroptera
Carnivora
Carnivora
Carnivora
Cetartiodactyla
Chiroptera
Perisodactyla
Pholidota
Macroscelidea
Chrysochloridae
Proboscidea
Sirenia
Hyracoidea
Tenrecidae
Macroscelidea
Tubulidentata
Pilosa
Cingulata
Pilosa
Scandentia
Glires
Monotremata
Marsupialia
Insectivora
Primates
Primates
Glires
Insectivora
Euarchontoglires
Laurasiatheria
Xenarthra
Afrotheria
Outgroup to Placentals

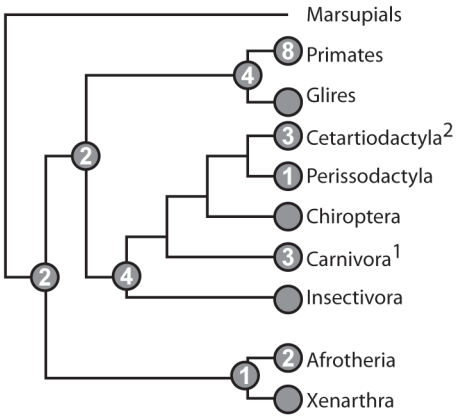
(A) Gene Coverage Across Taxa

	13	12	11	10	9	8	7	6	5	4	3	2
ATP6	25	34	45	54	73	79	103	145	189	204	230	253
ATP8	25	36	46	56	77	84	110	154	198	215	245	281
CO1	25	35	45	51	56	59	81	125	168	180	184	187
CO2	25	36	46	52	57	59	81	125	169	181	188	217
CO3	25	35	46	50	55	58	80	124	170	181	185	189
CYTB	25	32	35	44	61	83	106	137	143	154	180	267
ND1	25	36	47	58	82	103	119	126	127	127	128	129
ND2	25	36	46	57	81	106	124	129	129	131	135	152
ND3	25	36	45	56	77	99	113	115	115	133	138	141
ND4	25	36	47	58	83	107	123	130	132	153	158	163
ND4L	25	36	47	59	84	109	144	162	165	188	201	246
ND5	25	35	44	56	81	103	119	125	126	144	145	149
ND6	25	34	39	47	56	74	86	92	93	93	93	94
SM	25	36	47	59	84	109	147	197	244			

(B) % Phylogenetic Conflict

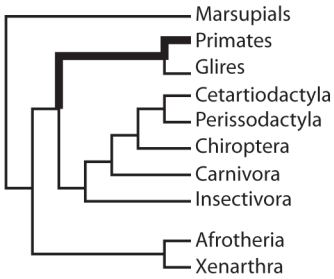

(A) ATP6
(B) ATP8
(C) CO1
(D) CO2
(E) CO3
(F) CYTB
(G) ND1
(H) ND2
(I) ND3
(J) ND4
(K) ND4L
(L) ND5
(M) ND6
(N) SM
Site Categories

(A) Summary of nodes tested

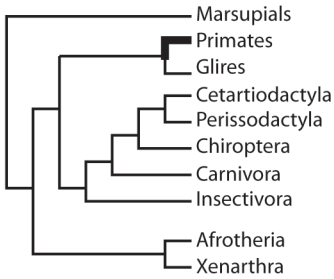


Node Tested	mtGene providing < 10% conflict
Placental mammal	ND4, ND5
Boreoeutheria	ND4, ND5
Euarchontoglires	ATP6, CO3, ND4, ND5
Laurasiatheria	CYTB, ND2, ND4,ND5
Atalantogenta	ND2
Primates	ATP6, CO1, CO2, CYTB, ND1, ND2, ND4, ND5
Cetartiodactyla	CYTB, ND2, ND5
Perissodactyla	CYTB
Carnivora	CYTB, ND2, ND5
Afrotheria	ND1, ND2

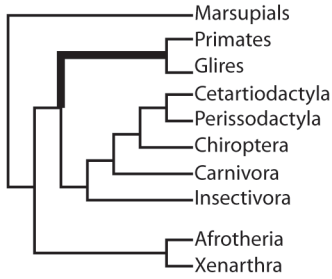
(B) ATP6



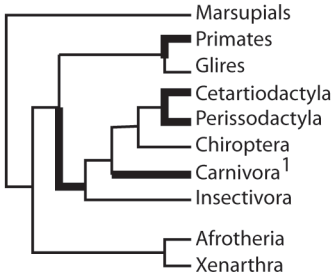
(C) CO1, CO2



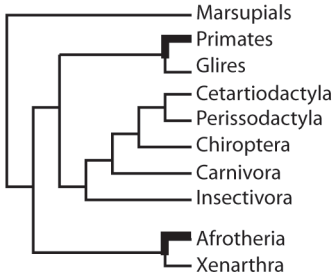
(D) CO3



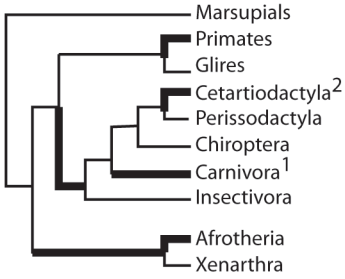
(E) CYTB



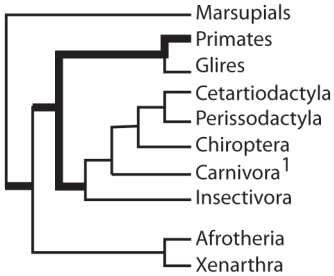
(F) ND1



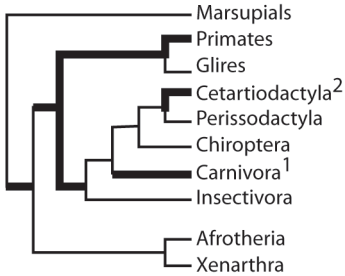
(G) ND2



(H) ND4



(I) ND5



(J) SM